*JUST BECAUSE IT'S RELIABLE
DOESN'T MEAN IT'S BELIEVABLE:
A Commentary on Two Articles by Birkimer and Brown*

THOMAS R. KRATOCHWILL

THE UNIVERSITY OF ARIZONA

The 1977 series of articles appearing in the *Journal of Applied Behavior Analysis* (Baer, 1977; Hartmann, 1977; Hopkins and Hermann, 1977; Kratochwill and Wetzel, 1977; Yelton, Wildman, and Erickson, 1977) prompted a somewhat unsettled state of affairs in the area of observer agreement in applied behavioral research. Although there was general consensus among these authors that the conventional percentage method of calculating "reliability" estimates did not adequately control for chance correspondence between observers' data, there was considerable disagreement among at least some of these authors over the most effective way to solve the problems.

Birkimer and Brown (1979a, 1979b) have done an admirable job in providing alternatives to some of the previous recommendations in reporting observer agreement data. Their procedures have much to recommend them and, if followed, will likely have an impact on future applied behavioral research.

Despite many positive aspects of both these reports, some words of caution are necessary. In the 1979b article, the authors demonstrate that although whole percentage agreement estimates are adequate as reliability measures (with due consideration for their limitations), tables developed to determine if obtained disagreements are unlikely due to chance are more easily employed. Alternatively, a rule developed in their

article makes the use of tables unnecessary in most cases. However, in these procedures, there is the potential hazard that some researchers will confuse statistical significance with meaningfulness. Observers should agree on the frequency (or other suitable measure) of the target behavior by some meaningful criteria—statistical decision rules can add to the overall decision rule as long as meaningful criteria of agreement are also applied. But there is the potential trap of increasing the total number of observations (N), thereby reaching statistical significance with no meaningful agreement between observers. Such a strategy would indeed be a case of the tail wagging the dog, a point Michael (1974) made when discussing the pitfalls of statistical significance tests. Thus the "easier ways" percentage measures are just as susceptible to this hazard as are other statistics employed in applied behavioral research.

There are more serious issues to consider in the 1979a article. In this paper the authors recommend a graphical procedure which summarizes obtained and chance reliability data that purportedly helps assess the believability of experimental effects. It may be important to consider the merit of these recommendations within the context of a distinction made by Kratochwill and Wetzel (1977) between the researchers (individuals conducting the research) and the research consumers (individuals reading the published experimental report). Birkimer and Brown (1979a) essentially recommend the procedures for the researcher *and* research con-

Reprints may be obtained from Thomas R. Kratochwill, Department of Educational Psychology, University of Arizona, Tucson, Arizona 85721.

sumer. I see the major limitations occurring when the graphical procedures are placed in published experimental reports, that is, presented for the research consumer. Several considerations emerge.

First, I do not believe that the authors have adequately demonstrated that their procedures are superior to various statistical measures used to assess observer agreement. The notion that such measures as correlationlike statistics are far removed from the basic data they summarize and therefore purportedly run counter to the behavior analysis tradition is not a sufficient reason to reject them. Indeed, the proposal that observer reliability data be plotted graphically is quite removed from the behavior analysis tradition which has relied on statistical judgmental aids in reporting observer data since its inception. Unfortunately, the statistical properties of the common percentage agreement method have not been elucidated in behavior analysis research until recently. Proposed solutions to this problem, for example, the use of correlationlike statistics (Hartmann, 1977), reporting tables of chance agreement (Hopkins and Hermann, 1977), and procedures for combining occurrence and nonoccurrence agreement scores (Harris and Lahey, 1978), have remained within the tradition of behavior analysis (and psychology in general) because reliability data are generally presented independent of the basic data, whether reported in summary statistic fashion from group research or in the graphical plot over successive phases in within-subject designs. As I have argued previously, statistical indices have promoted ease of judgment for the research consumer (Kratowill and Wetzel, 1977). Given that an appropriate statistic is reported (and establishing that other methodological concerns have been addressed), the research consumer can simply read the experimental report and find that agreement was assessed appropriately, various methodological factors were controlled, and get on to viewing the experimental data in the context of results and implications drawn. What Birkimer and Brown

(1979a) propose is that observer agreement data be integrated into the basic data that are used to draw conclusions from the research. This may be adding complexity where we do not need it.

My second concern with their recommendations has to do with the implications of graphically reporting observer reliability data on the "believability" of experimental effects. Presumably, this concern faces both researcher and research consumer. The recommendation to graphically portray data from more than one observer has been presented from time to time (e.g., Hawkins and Dotson, 1975; Morris, Rosen, and Clinton, Note 1). In Kratochwill and Wetzel (1977), we reviewed the recommendations of Hawkins and Dotson (1975) who argued that calibrating observer data should be plotted along with data from the primary observer in the research report. Basically, *with some important additions*, Birkimer and Brown (1979a) recommend this procedure. Specifically, the authors recommend that, in addition to plotting the calibrating observer's data for each reliability check, researchers should present disagreement percentages on graphs of the primary observer's data, centered halfway between the two observers' reported rates of target behavior. Also, researchers are advised to graph the limits of chance disagreement ranges for each reliability check. This provides all available information on chance agreement and permits determination of whether or not "true" observer agreement has been established. A presumed advantage of graphical representation is that such a strategy will help assess the believability of experimental effects for researchers and research consumers. The authors note:

For experimental effects to be believable they must be substantial enough to produce no overlap between disagreement ranges resulting from reliability checks taken before and after the claimed experimental effect occurred. If overlap exists between these disagreement ranges, the apparent

change in target behavior is not great enough, given the degree of observer agreement, to permit certainty that an effect of treatment has been shown (Birkimer and Brown, 1979a, p. 530).

The likelihood of detecting differences among conditions is a function of the reliability of the measures used. Thus, if there is overlap, the claim for believability is challenged by unreliability. However, the issue is not as straightforward as one might assume. There is a danger that individuals may treat these ranges as traditional confidence intervals and assume that because there is no overlap between disagreement ranges resulting from the reliability checks taken before and after the intervention, the effects are now believable, that is, the effects have been demonstrated. In actuality, all that has been demonstrated is that some pattern of data has occurred which is greater than the disagreement range between the two observers and which cannot be attributed to chance variations of observations. Finding good correspondence between observers does not address the issue of whether the effect from one phase to the next is a real or "chance" effect. Thus, when there is no overlap and chance variations of observations are not accounting for an effect, experimental outcomes should be evaluated by criteria independent of the observers' data. These criteria have been reviewed at length elsewhere (Jones, Vaught, and Weinrott, 1977; Kazdin, 1976, 1977; Kratochwill, 1978) and need not be discussed here.

Third, consider the alternative type of situation where there is overlap among *some* of the data points in the disagreement range. The authors suggest that when disagreement ranges overlap between conditions, a reliable change has not been demonstrated. However, when a target response is repeatedly assessed over time, we evaluate the phase changes in a within-subject design *on the basis of the entire series*, not on the basis of one data point within each phase. This issue raises the question of how many

observer checks should be scheduled and subsequently reported to ensure that observers are consistent in showing no overlap across the dependent measures used in the study. Presumably there will usually be fewer checks than total data points due to the cost and practical difficulties inherent in observational research in applied settings.

A fourth issue that must be considered is related to how much "overlap" will be tolerated on the disagreement range. In their hypothetical example presented in Figure 2 (Panel B) we have the rather idealized situation in which there is overlap among all four phase checks. But the situation will arise when some percentage of scores overlap out of some total number taken over the course of the study. (Even more difficult to evaluate is the situation in which some scores overlap to different degrees.) Of course, this situation occurs when statistical indices are used and the agreement score drops below an acceptable statistical criterion.

A final and perhaps the greatest potential problem with the graphical reliability procedures is increased complexity in determining if an experimental effect has occurred. Complexity will be high relative to present forms of analysis because more material will be placed on the graph. Graphical methods of analysis have already been chided for their unreliability in drawing conclusions about treatment effects (*cf.* Jones *et al.*, 1977; Kratochwill, 1978). Plotting data from both observers, as well as disagreement and chance ranges (and the many other procedures that frequently complement a graphical analysis), may further reduce the value of this method for researchers.

But, there are also some potential benefits of the graphical reliability procedures for researchers who want to use them during the experiment. The graphical reporting methods complement other reliability estimates and may help clarify for researchers special correlationlike estimates, tables of chance agreement, and even the procedures discussed by Harris and Lahey (1978). Note that Birkimer and Brown recom-

mend precise criteria for determining that obtained agreement is unlikely the result of chance agreement in their 1979*b* article—procedures that replace the graphed disagreement range recommended in the 1979*a* article. Presumably one might still compare the statistical and the graphical methods as an additional check.

Graphical procedures have several advantages relative to other techniques for handling chance agreement (or disagreement). As Birkimer and Brown (1979*a*) note, a graphic representation of disagreements is likely much simpler to present and understand than tables of obtained and chance S-I, U-I, and I-I agreement scores. Plotting the calibrating observer's data may allow researchers to detect critical behaviors that might otherwise go unnoticed, maintain contact with the data, evaluate absolute differences between/among observers, and relatively little sophistication in statistics is required in the evaluation of observer agreement (Kratowchwill and Wetzel, 1977).

Thus, graphical procedures may complement other strategies that also serve to make credible the basic data of the research report. However, even then, the primary benefits of graphical reporting occur prior to the publication of the experimental report. When data from two observers reveal an unacceptable estimate, which corresponds to statistical indices as well, the experiment should be reevaluated. The researcher should undertake the task of redesigning the observational technology, the nature of the intervention, or both.

In conclusion, I believe that readers of experimental reports are generally interested in knowing if the basic data were reliable through the most simple and straightforward judgmental aid. A major tradition in applied behavioral research, and behavior therapy research in general, has been to refer to a statistical judgmental aid for observer agreement data, graph the data from a primary observer, and draw conclusions about intervention effects from these data. This appears to be a convention worth preserving for the research consumer.

REFERENCE NOTE

1. Morris, E. K., Rosen, H. S., and Clinton, L. P. *Reliability considerations in single-subject research*. Paper presented at the meeting of the Midwestern Association for Behavior Analysis, Chicago, Illinois, May 1975.

REFERENCES

- Baer, D. M. Reviewer's comments: Just because it's reliable doesn't mean you can use it. *Journal of Applied Behavior Analysis*, 1977, **10**, 117-119.
- Birkimer, J. C. and Brown, J. H. A graphical judgmental aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis*, 1979, **12**, 523-533. (a)
- Birkimer, J. C. and Brown, J. H. Back to basics: Percentage agreement measures are adequate, but there are easier ways. *Journal of Applied Behavior Analysis*, 1979, **12**, 535-543. (b)
- Harris, F. C. and Lahey, B. B. A method for combining occurrence and nonoccurrence interobserver agreement scores. *Journal of Applied Behavior Analysis*, 1978, **11**, 523-527.
- Hartmann, D. P. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 1977, **10**, 103-116.
- Hawkins, R. P. and Dotson, V. A. Reliability scores that delude: An Alice in Wonderland Trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp and G. Semb (Eds), *Behavior analysis: Areas of research and application*. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.
- Hopkins, B. L. and Hermann, J. A. Evaluating interobserver agreement of interval data. *Journal of Applied Behavior Analysis*, 1977, **10**, 121-126.
- Jones, R. R., Vaught, R. S., and Weinrott, M. R. Time-series analysis in operant research. *Journal of Applied Behavior Analysis*, 1977, **10**, 151-166.
- Kazdin, A. E. Statistical analysis for single-case experimental design. In M. Hersen and D. Barlow, *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon, 1976.
- Kazdin, A. E. Assessing the clinical or applied importance of behavioral change through social validation. *Behavior Modification*, 1977, **1**, 427-452.
- Kratowchwill, T. R. (Ed) *Single subject research: Strategies for evaluating change*. New York: Academic Press, 1978.
- Kratowchwill, T. R. and Wetzel, R. J. Observer agreement, credibility, and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis*, 1977, **10**, 133-139.

- Michael, J. Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 1974, **7**, 647-653.
- Yelton, A. R., Wildman, B. G., and Erickson, M. T. A probability-based formula for calculating observer agreement. *Journal of Applied Behavior Analysis*, 1977, **10**, 127-131.
- Received 29 May 1979.
(Final Acceptance 9 July 1979.)